CollegeBoard

Research Report
No. 2004-1

New SAT® Writing Prompt Study: Analyses of Group Impact and Reliability

Hunter Breland
Melvin Kubota
Kristine Nickerson
Catherine Trapani
Michael Walker

connect to college success
www.collegeboard.com

College Board Research Report No. 2004-1
ETS RR-04-03

New SAT® Writing Prompt Study: Analyses of Group Impact and Reliability

Hunter Breland, Melvin Kubota,
Kristine Nickerson, Catherine Trapani,
and Michael Walker

College Entrance Examination Board, New York, 2004

Hunter Breland is a principal research scientist at HunterBreland.com.
Melvin Kubota is an assessment specialist III at Educational Testing Service.
Kristine Nickerson is a program administrator at Educational Testing Service.
Catherine Trapani is an associate research data analyst at Educational Testing Service.
Michael Walker is a measurement statistician–lead at Educational Testing Service.

The College Board: Expanding College Opportunity

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 4,500 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #30481024) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Acknowledgments

This project was funded by the College Board. In addition to the authors, a number of staff members at Educational Testing Service assisted with the data collection, administration, and report preparation. Betty Lou Estelle, Amy Dabbs, and Regina Mercadante provided administrative assistance; Jennifer Zanna and Patti Baron conducted statistical analyses associated with creating the data file; Nancy Glazer helped prepare the new persuasive prompts and select the training papers for the essay reading, Barbara Hames provided editorial assistance, and Christin Noble created the test booklet design. Rose Ann Morgan and Marilyn Collins led the essay reading. Yong-Won Lee provided useful statistical advice. Helpful reviews of report drafts were provided by Brent Bridgeman, Ida Lawrence, Yong-Won Lee, Larry Stricker, and Dan Eignor.

Contents

Abstract

This study investigated the impact on ethnic, language, and gender groups of a new kind of essay prompt type intended for use with the new SAT®. The study also generated estimates of the reliability of scores obtained using the prompts examined. To examine the impact of a new prompt type, random samples of eleventh-grade students in 49 participating high schools were administered writing tests using four different prompts, two of an old type and two of a new type. To obtain estimates of the reliability of scores for the old and new types of prompts, schools were asked to participate in a second round of testing to occur four months after the initial testing. Results of the impact analyses revealed no significant prompt type effects for ethnic, gender, or language groups, although there were significant differences in mean scores for ethnic and gender groups for all prompts. The score reliability estimates obtained were similar to those obtained in previous studies.

Keywords: writing prompts, ethnicity, language, gender, reliability

Introduction

During planning for the implementation of the new SAT, scheduled for the year 2005, a number of discussions were held to determine what kind of prompt should be used for the writing assessment. A decision on prompt type was of special interest because, for the first time on the SAT, very large numbers of students would be writing essays in response to the prompts. Currently, SAT essay test administrations are limited to those conducted for the SAT II: Writing Subject Test, a test most often required by the most selective colleges and limited in volume. One possibility considered was to use the same type of writing prompt as used for the current SAT II: Writing Test. Another possibility suggested was to use what has been termed a "persuasive" prompt type. Such persuasive prompts, while encouraging the examinee to be as persuasive as possible in his or her response, also usually provide more detailed information to the examinee. The additional information provided requires slightly more time for reading the prompt and the instructions.

A primary consideration in making a decision on the prompt type was whether changing to a new prompt type would have a negative impact on any ethnic or gender group or on examinees for whom English is a second language. Although there has been no previous research on the comparative impact of prompt types on ethnic, gender, or language groups, there have been a number of studies of ethnic, gender, and language group differences for the prompts used in several different assessments.

For ethnic groups, three studies were conducted for the English Composition Test (ECT), a precursor of the SAT II: Writing Subject Test (Breland and Jones, 1982; Pomplun, Wright, Oleka, and Sudlow, 1992; Breland Bonner, and Kubota, 1995). Ethnic differences on the California State Universities and Colleges English Placement Test (EPT) were examined by Breland and Griswold (1982). At the graduate level, studies of ethnic differences on essay tests have been conducted by Bridgeman and McHale (1996) for the Graduate Management Admission Test (GMAT), by Schaeffer, Briel, and Fowles (2001) for the Graduate Record Examination (GRE), and by the American Association of Medical Colleges (1997) for the Medical College Admission Test (MCAT). The results of these studies show considerable differences in performance between white and African American, Asian American, and Hispanic examinees. African American/white essay performance differences in college-bound populations averaged between one-third and one-half of a standard deviation, while at the graduate level the difference was about two-thirds of a standard deviation. Hispanic/white differences in essay performance averaged between one-third and one-half of a standard deviation in college-bound populations and about one-half of a standard deviation in graduate populations. Asian American/white differences in essay performance averaged between one-third and one-half of a standard deviation in college-bound populations but higher (about three-fourths of a standard deviation) at the graduate level.

Differences in essay writing performance for gender groups vary somewhat depending on the population. For national random samples of students, the National Assessment of Educational Progress (NAEP) has observed differences in essay writing performance of about one-half of a standard deviation (favoring females) for samples of Grade 8, Grade 11, and Grade 12 students (NAEP, 1994). Engelhard, Gordon, and Gabrielson (1991) observed differences of about the same magnitude for Grade 8 students in Georgia. Gender differences for college-bound populations tend to be smaller, ranging from about one-tenth to one-third of a standard deviation, favoring females (Breland and Griswold, 1982; Breland and Jones, 1982; Bridgeman and Bonner, 1994; Pomplun et al., 1992). At the graduate level, gender differences are about the same as for college-bound populations (AAMC, 1997; Bridgeman and McHale, 1996; Schaeffer et al., 2001).

There have not been many studies of students for whom English is a second language. One study of lan-

guage group differences found that Hispanic and Asian American ESL students performed about three-fourths of a standard deviation lower than white students (Pomplun et al., 1992) in a college-bound population, but there are some indications that the difference observed may vary with the population studied. That is, more selective populations may have larger essay performance differences than national random samples. One study of medical school applicants, for example, observed differences between white and Hispanic ESL students of two standard deviations (AAMC, 1997).

Because the proposed "persuasive" prompt type would provide more information, it was of interest to consider what research may have been conducted concerning the length of essay prompts. Ruth and Murphy (1988) summarize one study of "information load" conducted by Brossell (1986). In this study writing prompts with "low," "moderate," and "high" levels of information load were compared. One prompt with a low level of information load consisted of only four words, while two prompts with moderate and high information loads contained 29 and 107 words, respectively. Six different topics were used in the study, each with the three levels of information load. The prompts were administered randomly to 360 undergraduate education majors and scored by three different raters. No statistically significant differences in scores were observed across the three levels of information load. Despite these results, some readers of Brossell's study still believed that longer prompts tend to introduce problems. Hoetker, Brossell, and Ash (1981), for example, made the following comment concerning the longer prompts:

> First, such a scenario introduces into the testing situation all of the problems of varying individual interpretations and responses that are associated with the reading of any work of fiction. Second, the sheer amount of language that students must process is increased. Opportunities for confusion, misinterpretation, and creative misreadings are proportionately increased. Third, the more language and information students are given the more difficult it seems to be for them to get beyond the language of the topic to discover what they may themselves have to say, so that examiners find themselves receiving not 'original responses,' but their 'own prose back in copy speech.'

Given such beliefs, a study of the effects of different prompt types seemed to be important. Especially important would be to examine prompt types for differential effects within ethnic groups, which was the objective of the present study.

Study Design

The principal constraint in the study design was the study schedule, which called for administration of the prompts to be studied in November 2002, the scoring of the responses in December 2002, and the reporting of preliminary results in January 2003. Because of reader availability, competing scoring requirements, and other factors, it was determined that the scoring would have to be conducted in a single day. In a single day, it was estimated that about 4,800 readings could be conducted, or two readings of each of 2,400 essays.

Although it was not considered in the original study design, a reliability study was designed following the November data collection. A number of schools agreed to have their students write a second essay on a different topic in March 2003. The second essays were then scored using the same procedures followed in December. Comparisons of the scores received by students in the two administrations formed the basis for the reliability analyses.

Sampling

Because sampling was required for four ethnic groups (African American, Asian American, Hispanic, and white), the time constraint meant that a total of approximately 600 students from each group could be sampled. The only remaining question was how many different prompts and types of prompts could be studied.

One study design considered was to use three different topics within each of two different prompt types. The first prompt type would be that used for the SAT II: Writing Subject Test and the second an elaborated version of the SAT II prompt that encouraged persuasive writing and provided more information to examinees. Thus, within each ethnic group sampled, there would be six different treatments of which three would be based on SAT II type prompts and three based on persuasive type prompts. This design would have six treatments X four ethnic groups and 24 individual cells. With a total of 2,400 students sampled, such a design would have 100 students within each cell. Power analysis (Cohen, 1988) and considerations of possibly poor responses from schools, student motivational problems that could produce unscorable responses, and other considerations led to a decision to reject this initial design model.

An alternate design using two prompts within each of two prompt types was ultimately chosen, and this design is shown in Table 1. As indicated, the alternate design would allow for approximately 150 students in each cell, instead of only 100 as for the initial design. The sampling of language and gender groups would

TABLE 1

Study Design

Ethnic Sampling

	SAT® II Writing Prompt		*Persuasive Writing Prompt*	
Group	*A1*	*A2*	*B1*	*B2*
Asian American	150	150	150	150
African American	150	150	150	150
Hispanic	150	150	150	150
White	150	150	150	150
Total	600	600	600	600

Grand total of students = 2,400
Number of readings = 4,800

Language Group Sampling

	SAT II Writing Prompt		*Persuasive Writing Prompt*	
Group	*A1*	*A2*	*B1*	*B2*
ESL	180	180	180	180
Non-ESL	420	420	420	420
Total	600	600	600	600

Gender Group Sampling

	SAT II Writing Prompt		*Persuasive Writing Prompt*	
Group	*A1*	*A2*	*B1*	*B2*
Female	300	300	300	300
Male	300	300	300	300
Total	600	600	600	600

depend on the sampling of ethnic groups, but it was estimated that about one-third of sampled students would be students for whom English was a second language as indicated in the middle of Table 1. It was estimated that genders would be approximately equal, as indicated at the bottom of Table 1.

Instruments

The instruments used for the study were two regular SAT II: Writing Subject Test prompts and two modifications of these prompts that would encourage persuasive writing and provide more information to the examinee. The first SAT II prompt (coded A1) was on the topic of failure. The modification of prompt A1 (coded B1) was also on the topic of failure but provided more information and encouraged persuasive writing. The second SAT II prompt (coded A2) was on the topic of happiness. The modification of prompt A2 (coded B2) was also on the topic of happiness but provided more information and encouraged persuasive writing. The essay prompts used in the study are included in this report as Appendix A.

Scoring of responses was conducted by two groups of readers working independently, with each reader assigning a score of 1 (low) to 6 (high). The two reader scores for each essay were then summed. In the event of a discrepancy of two or more score points, a third reader was used to resolve the discrepancy.

Data Collection

Data collection was based on school information obtained from the 2001 PSAT/NMSQT® data file, which indicated the ethnic distribution of students in schools. A total of 500 schools were selected such that the aggregate of them would contain an equal balance of African American, Asian American, Hispanic, and white students. These selected schools were sent invitations to participate in the study (see Appendix B for the letters sent) and asked to provide information about the number of eleventh-grade students who would probably participate. As an incentive, schools were told that Educational Testing Service would score their student responses, and the scored essays would be returned to them. Additionally, schools were promised statistical information comparing their students' scores with the scores in our national sample of students.

A total of 130 schools responded and indicated that they would like to participate. A number of schools wanted to include their entire eleventh-grade class, and many wanted to test a large proportion of their eleventh-grade class. With a study limitation of approximately 2,400 students, it was necessary to reduce the number of schools and to control the numbers of students within each school. Fifty schools were selected for participation, with a cap of 60 students allowed per school. Some schools decided not to participate with these limitations, and replacement schools were selected. A total of 49 schools ultimately participated in the study. A list of the participating schools is included in this report as Appendix C.

Test booklets were designed such that student identification information, school identification information, and information on students' genders, ethnic identification, and language experiences were obtained from the participating students. Teachers in the schools who participated in the study were instructed to open the test booklet packets and to distribute test booklets randomly to students in the participating classes by selecting booklets sequentially from the top of each packet.

Data Analyses

Two types of data analyses were conducted: (1) analyses of mean differences for different ethnic, gender, and language groups, and (2) analyses for the estimation of the reliabilities of the instruments used.

Analyses of Mean Differences

The analyses of mean differences were conducted using analysis of variance. Two types of ANOVA models were used. In one ANOVA model, specific prompt topics were nested within prompt types to examine overall effects of prompt and topic on different groups. In a second type of ANOVA model, prompts were analyzed individually without regard to prompt type. In both types of models, both main effects and interactions were analyzed.

Reliability Estimates

Reliabilities were estimated using multiple methods. First, Pearson correlations were computed between scores obtained by students on different occasions on two different forms of each prompt type. Second, coefficient Alpha was computed for the two scores obtained on the two different occasions and then stepped down to one essay using the Spearman-Brown formula. Third, an intraclass correlation was computed using the two scores from the two different forms using methods outlined in Shrout and Fleiss (1979) and Winer (1962). The intraclass correlations assumed that each subject was rated by multiple raters, that raters were randomly assigned to subjects, and that all subjects had the same number of raters. Finally, generalizability coefficients were computed using methods outlined in Brennan (1992).

Results

Table 2 describes the data obtained in the study by ethnic group and prompt code. The number of cases obtained for analysis varied only slightly across different ethnic groups and prompt codes, with the largest number of cases (161) being obtained for white students who responded to the A1 prompt. The fewest cases (139) were obtained for African American students who responded to the B1 prompt. When students for whom English was not their best language were excluded, the cases available for analysis were reduced somewhat but not to a degree that would severely affect analyses. With this exclusion, the largest number of cases available for analysis (158) occurred for white students who wrote on the A1 prompt, as for the total sample. The smallest cells (127 cases) were for Asian American students who responded to the A1 and B1 prompts and for Hispanic students who responded to the A1 prompt. The highest observed mean essay score (7.22) was for white students who responded to prompt B2, and the lowest observed mean score (5.45) was for African American students who responded to prompt A2. The ranges of scores obtained varied some for different ethnic groups, with the smallest range (2 to 10) occurring for African American students who responded to prompts A1, B1, and B2, and for Hispanic students who responded to prompts A2 and B2.

Table 2 describes the data obtained by prompt code and gender and language groups. More female than male cases were obtained for all four prompts. English was the first language learned by most respondents (about 70 percent), and English was the best language for almost all respondents (over 90 percent). The mean scores obtained by females appear to be higher than those for males for all prompts, while the mean scores for English First Language and English Not First Language groups do not appear to differ much. The mean scores for English Best Language (EBL) and English Not Best Language (ENBL) groups are somewhat different, as would be expected. The range of scores obtained was also lowest for ENBL students.

TABLE 2

Total Essay Score Data Description by Ethnic Group and Prompt Code

Group/Prompt	*N*	*Mean*	*S.D.*	*Min.*	*Max.*
Asian American					
A1	143	6.92	2.02	2.0	11.0
A2	152	7.15	2.07	2.0	12.0
B1	152	7.11	2.21	2.0	12.0
B2	151	6.91	2.41	2.0	12.0
African American					
A1	143	6.03	2.00	2.0	10.0
A2	144	5.47	2.02	2.0	11.0
B1	139	5.69	1.85	2.0	10.0
B2	152	5.53	1.99	2.0	10.0
Hispanic					
A1	147	6.22	2.03	2.0	11.0
A2	160	6.06	1.90	2.0	10.0
B1	162	6.23	2.03	2.0	11.0
B2	152	6.09	2.05	2.0	10.0
White					
A1	161	6.86	1.95	2.0	11.0
A2	143	7.01	2.08	2.0	12.0
B1	148	6.93	2.17	2.0	11.0
B2	144	7.22	2.00	2.0	12.0

TABLE 3

Total Essay Score Data Description by Gender, Language Group, and Prompt Code

	Statistic				
Group/Prompt	*N*	*Mean*	*S.D.*	*Min.*	*Max.*
Female					
A1	324	6.76	1.95	2.0	11.0
A2	352	6.75	2.01	2.0	12.0
B1	346	6.68	1.98	2.0	11.0
B2	351	6.73	2.12	2.0	12.0
Male					
A1	295	6.22	2.10	2.0	11.0
A2	270	6.08	2.30	2.0	12.0
B1	275	6.27	2.32	2.0	12.0
B2	265	6.08	2.28	2.0	12.0
English First					
A1	453	6.51	2.05	2.0	11.0
A2	427	6.46	2.16	2.0	12.0
B1	432	6.50	2.16	2.0	12.0
B2	436	6.45	2.18	2.0	12.0
English Not First					
A1	166	6.48	2.02	2.0	11.0
A2	193	6.42	2.17	2.0	12.0
B1	187	6.53	2.11	2.0	11.0
B2	180	6.45	2.30	2.0	12.0
English Best					
A1	575	6.59	1.99	2.0	11.0
A2	554	6.58	2.16	2.0	12.0
B1	557	6.59	2.13	2.0	12.0
B2	573	6.53	2.19	2.0	12.0
English Not Best					
A1	35	5.26	2.27	2.0	10.0
A2	64	5.25	1.81	2.0	10.0
B1	61	5.82	2.10	2.0	10.0
B2	41	5.37	2.33	2.0	11.0

Ethnic Group Differences

Figure 1 depicts the differences in mean scores obtained for the four ethnic groups on the SAT II and persuasive type prompts. The tick marks at the top of each bar in the graph represent the 95 percent confidence interval around the mean score for each group and prompt type. For all ethnic groups, there were small differences in the means obtained, but none of these differences falls outside of the 95 percent confidence interval, and none is statistically significant at the .05 level. The largest observed difference occurred for African American students, with the SAT II type prompt yielding slightly higher mean scores. This difference for African American students was not statistically significant, however. Additional details of this analysis are given in Appendix D.

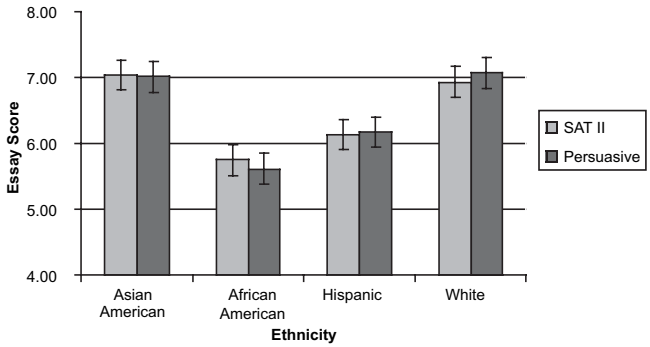

Figure 1. Mean essay scores by prompt type and ethnicity.

Figure 2 presents the results by prompt code, without collapsing across the two prompts within each type. This figure illustrates not only the difference between the two prompts (SAT II and persuasive), but also the variability between the two prompts within type. Additionally, the analyses represented in Figure 2 were conducted both for all students (Figure 2a) and for students who reported that English was their best language (Figure 2b). Figure 2 indicates that the mean essay scores obtained for four ethnic groups on the four prompts, while different, did not vary appreciably across prompts within each group. The analysis of variance results also showed that there

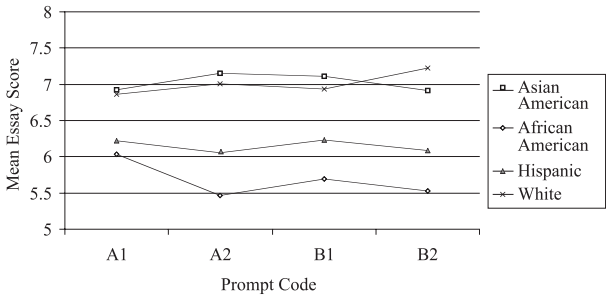

Figure 2a. All students.

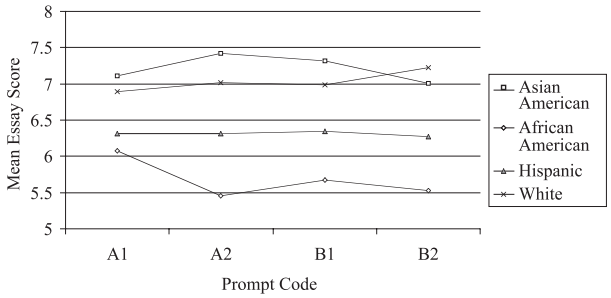

Figure 2b. English best language students.

Figure 2. Ethnic performance on different essay prompts.

was no statistically significant interaction between ethnic group and prompt code. The only statistically significant difference across prompts occurred between prompts A1 and the remaining three prompts for African American students. The reason for this difference is unknown, but it is not likely the result of sampling because the same pattern occurred for both African American female and male students. This finding is more interesting when it is recalled that Prompts A1 and B1 covered the same topic using different prompt types. A contrast of Figures 2a and 2b shows that Asian American student mean scores increased a little, making a greater separation between white and Asian American students. Hispanic student mean scores also increased a little, and the means across prompts were almost perfectly flat. Further details of the ethnic analyses are given in Appendix D.

Gender Group Differences

Figure 3 depicts gender differences for the four prompts and indicates differences consistent with those observed in previous studies. These gender differences are statistically significant at the .05 level, with an effect size ranging from .19 (for prompt B1) to .31 (for prompt A2). An effect size of .20 is considered "small," in the scheme devised by Cohen (1988), and .50 to be "medium." Further details of the gender analyses are given in Appendix D.

Language Group Differences

Figure 4 depicts graphically differences in mean essay scores for English First Language students and English Not First Language students. The scale for the graph is the same as that used for the ethnic and gender comparisons. There are no statistically significant differences in mean essay scores for any of the four prompts, and there are no statistically significant differences in mean essay scores across prompts. This result suggests that the students sampled had been attending U.S. schools for some time such that, even though their first

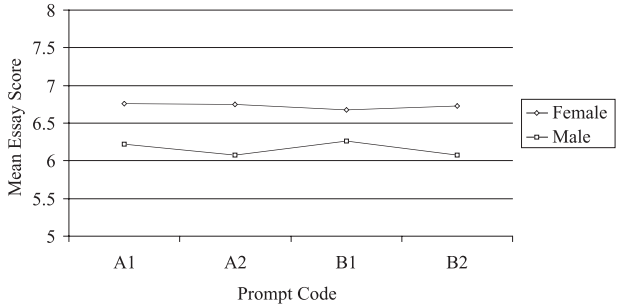

Figure 3. Gender performance on different essay prompts.

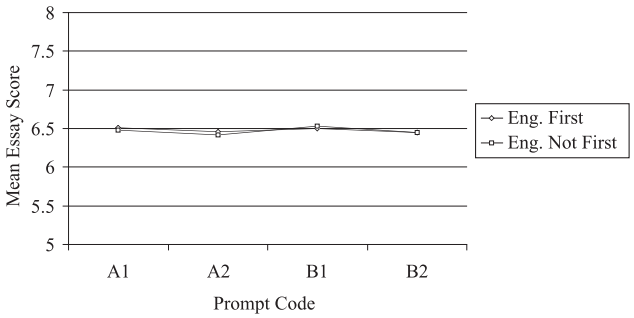

Figure 4. Language group performance on different essay prompts.

language was not English, they had attained by Grade 11 a good command of English and were thus able to perform as well on the essay writing tasks as students whose first language was English.

Although the numbers of cases available for analysis were small for students for whom English was not their best language, comparisons were made between English Best Language (EBL) students and English Not Best Language (ENBL) students. For all four prompts, EBL students performed significantly better than ENBL students ($p < .05$). The effect sizes ranged from .36 (for prompt B1) to .66 (for prompt A1). This makes an interesting contrast between ENBL students and African American students. While African American students performed best on Prompt A1, ENBL students performed worst on this prompt. Further details of the language group analyses are given in Appendix D.

Reliability Study

Table 4 gives reliability estimates for the SAT II and persuasive prompts used in the current study. Table 4 shows that 138 students responded to prompt A1 in November, and that these same students responded to prompt A2 in March; 131 students responded to prompt A2 in November, and these same students responded to prompt A1 in March; 128 students responded to prompt B1 in November, and these same 128 students responded to prompt B2 in March; 131 students responded to prompt B2 in November, and these same 131 students responded to prompt B1 in March. The mean essay scores in Table 4 show that students did gain in their writing performance between November and March. These score gains represent effect sizes ranging from .16 to .29, and all are statistically significant differences ($p < .05$) with the exception of the lowest gain (.16).

Reliabilities for the essay assessments were estimated using a variety of methods, as indicated in Table 4. Data for the two different orders of administration were first

TABLE 4

Reliability Estimation

	Prompt Type							
	SAT II Prompts				*Persuasive Prompts*			
	A1		*A2*		*B1*		*B2*	
Statistic	*NOV*	*MAR*	*NOV*	*MAR*	*NOV*	*MAR*	*NOV*	*MAR*
N	138	131	131	138	128	131	131	128
Mean	6.67	7.16	6.79	6.97	6.55	7.06	6.70	7.12
S.D.	1.97	2.01	2.05	2.18	1.86	2.21	2.16	2.05
Min.	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
Max.	11.00	12.00	12.00	12.00	11.00	12.00	12.00	12.00

	Reliability Indexes			
Prompt Type	*Pearson Correlation*	*Coefficient Alpha*	*Intraclass Correlation*	*Generalizability Coefficient*
SAT II	.59	.60	.58	.60
Persuasive	.56	.55	.54	.56

combined such that 269 students who wrote on both the A1 and A2 prompts and 259 students who wrote on both the B1 and B2 prompts were available for analysis. Table 4 shows that the Pearson correlation between the essays written for the SAT II prompts (A1 and A2) was .59 and that between the persuasive prompts (B1 and B2) was .56. Coefficient Alpha produced similar results (.60 and .55) after stepping down to one essay using the Spearman-Brown formula. The Intraclass Correlation for the SAT II prompts was .58 and that for the persuasive prompts .54. The generalizability coefficients were .60 for the SAT II prompts and .56 for the persuasive prompts.

Table 5 indicates what students might expect if they were to repeat an essay test of writing skill (of either an SAT II or a persuasive type) after a four-month period in which they were enrolled in high school and studying English composition. For example, for students who scored in the 10–12 range the first time they took the essay test, about half will have an increase or decrease of one point, about a third will have a decrease of 2 to 3 points, and about 10 percent will score four or more points less. For the average student who scored in the 6 to 7 range, well over half can expect to score within one point of their initial score, about one-fourth can expect an increase of 2 to 3 points, and about one-eighth can expect a decrease of 2 to 3 points. For students who scored very low on the initial assessment, most will either stay at about the same score or have an increase in score of 2 to 3 points. Only about one-fifth can expect to increase their score by 4 or more points. There are also, of course, floor and ceiling effects. Students who obtain the maximum possible score of 12 on the first testing can only obtain the same score or a lower score on retesting. Students who obtain the lowest possible score of 2 on the first testing can only obtain the same or a higher score on retesting.

TABLE 5

Probable Score Change from November to March for Essay Assessments of Writing Skill

	Percentage of Students with Gain or Loss after Taking an Essay Test of Writing Skill in November of their Junior Year and Again in March					
November Score	*Decreased 4 or more points*	*Decreased 2 to 3 points*	*Increased or decreased by 1 point*	*Increased by 2 to 3 points*	*Increased 4 or more points*	*Number of Test-Takers*
10–12	10	34	54	2		41
8–9	6	12	67	15	1	156
6–7	2	12	59	24	3	187
4–5		3	54	29	14	123
2–3			41	41	19	27

Discussion

Table 6 compares ethnic, gender, and language group differences (effect sizes) observed in the present study with differences observed in other studies. The comparisons are made with studies grouped by the type of populations involved in each of the studies. The NAEP (1994) data were obtained from a national random sample of high school students. The Engelhard et al. (1991) data were for a large sample of high school students who took a statewide assessment in Georgia. The statewide assessment was not administered to students in private high schools, however. Thus, both of these studies represented large numbers of high school students, but the Engelhard et al. sample probably excludes some of the best students in Georgia.

The Breland and Griswold (1982) study, which was of first-year students enrolled in California State Universities and Colleges (CSUC), represents a small step up in population selectivity. CSUC institutions are the least selective public postsecondary institutions in the state of California.

The Breland and Jones (1982), Pomplun et al. (1992), and Breland et al. (1995) studies were of college-bound students who had taken the English Composition Test (ECT) or the SAT II English essay test. Since the ECT and SAT II are required primarily by the most selective colleges and universities, these studies were of select groups of high school students.

The final group of studies in Table 6 were of graduate populations. Bridgeman and McHale (1996) studied students who had taken the Graduate Management Admission Test (GMAT), Schaeffer et al. students who had taken the Graduate Record Examination, and AAMC (1997) students who had taken the Medical College Admission Test (MCAT).

TABLE 6

Comparison of Ethnic, Gender, and Language Group Differences Observed in the Present Study with Differences Observed in Other Studies of Essay-Writing Skills

	Focal Group Impact (Effect Size)				
Study	*Asian American*	*African American*	*Hispanic*	*Female*	*ESL*
Current (SAT II Prompts)	-.04	.58	.38	-.29	.04
Current (Persuasive Prompts)	-.01	.64	.41	-.21	-.01
Oh and Walker (2003) SAT II	-.03	.45	.40	-.36	.16
Oh and Walker (2003) Persuasive	.02	.46	.42	-.36	.26
NAEP (1994)	-.02	.59	.35	-.54	
Engelhard et al. (1991)				-.38	
Breland and Griswold (1981)	.45	.81	.57	-.36	
Breland and Jones (1982)		.48	.48	-.16	
Pomplun et al. (1992)	.06	.37	.25	-.14	.74
Breland et al. (1995)	.36	.46	.34	-.34	
Bridgeman and McHale (1996)	.72	.71	.46	-.12	
Schaeffer et al. (2001)	.44	.76	.26	-.12	
AAMC (1997)	.00	.65	.32	-.13	

Notes:

(1) Effect sizes for ethnic focal group were computed as white mean minus Focal Group mean divided by average standard deviations.

(2) Effect size for Female was computed as Male minus Female divided by average standard deviation.

(3) Effect size for ESL was computed as English First Language mean minus English Not First Language mean divided by average standard deviation.

(4) The Asian American sample in the Pomplun et al. (1992) study was limited to students who reported that English was their first language.

Table 6 indicates that the group differences observed in the present study are similar to those that have been observed in other studies of essay writing performance. In all of these other studies, female students have outperformed male students. And in all of the studies, white students have performed better than African American or Hispanic students. There are some differences between the present study and other studies for Asian American and ESL students, but these differences may be because of differences in the populations studied. For example, the larger effect sizes obtained for Asian American students in the GMAT and GRE studies may be because many of these students are from foreign countries. And the effect size for ESL students observed in the Pomplun et al. (1992) study may be related to the selectivity of their sample. The present study, like the Oh and Walker (2003) study, was of a national sample of high school students who were not selected on the basis of ability. Although their first language may not have been English, most of these students had probably been in U.S. schools for most of their lives. In contrast, the Pomplun et al. study was of high school students applying for admission to selective colleges and universities. Thus, in the Pomplun et al. study, ESL students are being compared to very capable students. The ESL differences observed in the Oh and Walker study are slightly larger than those observed in the present study, but the effects are still small.

We did observe language differences between students who reported that English was their best language and students who reported that English was not their best language. These differences ranged from about one-third of a standard deviation to about two-thirds of a standard deviation for the different prompts examined. It was not possible to make precise comparisons for different prompts because the distribution of English Not Best Language students was not uniform across prompts.

Although differences in mean essay scores were observed for different ethnic, gender, and language groups in the present study, only one statistically significant difference was observed across prompt types within group. The one statistically significant group difference across prompts occurred in the African American group between prompt A1 and the other three prompts. It is not clear why the African American group performed better on prompt A1 than on the other three prompts. Because this anomaly occurred within both genders of African American students, it would not appear to be due to sampling error. Moreover, an examination of PSAT/NMSQT mean scores for the prompt/ethnic groups gave no indication that African American students who received prompt A1 were of higher ability.

The reliability estimates of the current study are similar to those made in previous studies. Wright (1992) obtained a reliability estimate of .58 for English Composition Test (ECT) essays based on a Pearson correlation. Schaeffer et al. (2001) obtained a reliability estimate of .62 for the Issue essay of the Graduate Record Examination Analytical Writing Assessment, which uses prompts similar to those used for the ECT and SAT II essay tests. The Schaeffer et al. study collected all data at one point in time, however, which might explain the slightly higher estimate.

Conclusion

The results of this study indicate that there should be no significant impact on any ethnic group of changing from an SAT II type writing prompt to a persuasive prompt of the type examined in this study. The one note of caution concerns the differences in performance between prompt A1 and the other prompts for African American

students, which would suggest an advantage of using the SAT II type prompt, at least for this topic. It would be important for future research to address this finding to determine whether it was an anomaly or a more robust occurrence. The results also indicate that there would be no significant impact on students for whom English is a second language if the writing prompt type were changed from an SAT II type prompt to a persuasive prompt of the type examined. The results of this study, as well as those of previous studies, indicate that female students can be expected to score significantly higher than male students on an essay assessment of writing skill.

The reliability estimates obtained in the current study, as well as those obtained in previous studies, indicate that essay assessments of writing skill are not as reliable as most traditional educational assessments. Consequently, students should expect that their scores may change appreciably if they take the same kind of test again within a few months. Students with high scores on their first test have a good probability of scoring lower on the second test. Most students who scored very low on the first test, however, can expect a higher score on the second test.

References

American Association of Medical Colleges (1997). *Characteristics of 1996 MCAT Examinees*. Washington, DC: AAMC.

Breland, H. M. (1977). *Group comparisons for the Test of Standard Written English*. College Board Research and Development Report RDR 77-78, No. 1 (Educational Testing Service Research Bulletin RB-77-15). Princeton, NJ: Educational Testing Service.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework* (College Board Report No. 99-3, GRE Board Research Report No. 96-12R, ETS RR No. 99-3). New York: College Board.

Breland, H. M., Camp, R., Jones, R.J., Rock, D., & Morris, M. (1987). *Assessing writing skill*. New York: College Board.

Breland, H. M. & Gaynor, J. (1979). A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement*, 16 (2), 119–128.

Breland, H. M. & Griswold, P. A. (1982). Use of a performance test as a criterion in a differential validity study. *Journal of Educational Psychology*, 74 (5), 713–721.

Breland, H. M. & Jones, R. J. (1982). *Perceptions of writing skill*. College Board Report No. 82-4 (ETS RR No. 82-47). New York: College Entrance Examination Board.

Breland, H. M., Bonner, M. W., & Kubota, M. (1995). *Factors in performance on brief, impromptu essay examinations*. College Board Report No. 95-4 (ETS RR No. 95-41). New York: College Entrance Examination Board.

Breland, H. M., Muraki, E., & Lee, Y. (2001). *Comparability of TOEFL CBT writing prompts*. Princeton, NJ: Educational Testing Service.

Brennan, R. (1992). *Elements of generalizability theory*. Iowa City: ACT.

Bridgeman, B. & Bonner, M. (1994). SAT-W as a predictor of grades in freshman English courses. Unpublished report. Princeton, NJ: Educational Testing Service.

Bridgeman, B. & Carlson, S. B. (1984). Survey of academic writing tasks. *Written Communication*, 1 (2), 247–279.

Bridgeman, B. & McHale, F. (1996). *Gender and ethnic group differences on the GMAT Analytical Writing Assessment* (ETS RR-96-2). Princeton, NJ: Educational Testing Service.

Brossell, G. (1986). Current research and unanswered questions in writing assessment. In K. Greenberg, H. Wiener, & R. Donovan (Eds.), *Writing assessment: Issues and strategies*. New York: Longman.

Coffman, W. C. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement*, 3(2), 151–156.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, second edition, Hillsdale, NJ: Erlbaum.

Engelhard, G., Gordon, B., & Gabrielson, S. (1991). The influences of mode of discourse, experiental demand, and gender on the quality of student writing. *Research in the Teaching of English*, 26 (3), 315–336.

Hoetker, J., Brossell, G., & Ash, B. (1981). *Creating essay examination topics*. Unpublished manuscript, Florida State University, Tallahassee.

National Assessment of Educational Progress (1994). *NAEP 1992 trends in academic progress*. Washington, DC: National Center for Education Statistics.

Oh, H. & Walker, M. (2003). The effects of essay placement and prompt type on performance on the New SAT. Draft Report. Princeton, NJ: Educational Testing Service.

Pomplun, M., Wright, D., Oleka, N., & Sudlow, M. (1992). An analysis of English composition essay prompts for differential performance. College Board Report No. 92-4. New York: College Entrance Examination Board.

Ruth, L. & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.

Schaeffer, G. A., Briel, J. B., & Fowles, M. E. (2001). *Psychometric evaluation of the new GRE Writing Assessment* (GRE Report No. 96-11P, ETS RR-01-08). Princeton, NJ: Educational Testing Service.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.

Winer, B. J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill.

Wright, D. (1992). Writing test field trial report (unpublished).

Appendix A: Essay Prompts

Prompt A1

ESSAY TOPIC

Time—25 minutes

Consider carefully the following statement and the assignment below it. Then plan and write your essay as directed.

Failure often contains the seeds of success.

Assignment: The statement above suggests that failure may be the source of success. In an essay, discuss the statement above, using an example (or examples) from literature, the arts, history, current events, politics, science and technology, or your experience or observation.

Prompt A2

ESSAY TOPIC

Time—25 minutes

Consider carefully the following statement and the assignment below it. Then plan and write your essay as directed.

The more you know, the happier you are.

Assignment: Decide whether you agree or disagree with the statement above. In an essay, support, challenge, or modify this statement, using an example (or examples) from literature, the arts, history, politics, science and technology, current events, or your experience or observation.

Prompt B1

ESSAY TOPIC

Time—25 minutes

Consider carefully the following excerpt and the assignment below it. Then plan and write an essay that explains your ideas as persuasively as possible. Keep in mind that the support you provide—both reasons and examples—will help make your view convincing to the reader.

> The principle is this: each failure leads us closer to deeper knowledge, to greater creativity in understanding old data, to new lines of inquiry. Thomas Edison experienced 10,000 failures before he succeeded in perfecting the lightbulb. When a friend of his remarked that 10,000 failures were a lot, Edison replied, "I didn't fail 10,000 times, I successfully eliminated 10,000 materials and combinations that didn't work."
>
> —Miles Brand, "Taking the Measure of Your Success"

Assignment: What is your view on the idea that it takes failure to achieve success? In an essay, support your position using an example (or examples) from literature, the arts, history, current events, politics, science and technology, or your experience or observation.

Prompt B2

ESSAY TOPIC

Time—25 minutes

Consider carefully the following excerpt and the assignment below it. Then plan and write an essay that explains your ideas as persuasively as possible. Keep in mind that the support you provide—both reasons and examples—will help make your view convincing to the reader.

> The well-known proverb 'Ignorance is bliss' suggests that people with knowledge of the world's complexities and its limitations are often unhappy, while their less-knowledgeable counterparts remain contented. But how accurate is this folk wisdom? A recent study showed that well-informed people were more likely to report feelings of well-being. In fact, more knowledge leads people to feel better about themselves and more satisfied with their lives.
>
> —adapted from Lee Sigelman, "Is Ignorance Bliss? A Reconsideration of the Folk Wisdom"

Assignment: What is your view on the idea that more knowledge makes people happier? In your essay, support your position using an example (or examples) from literature, the arts, history, science and technology, politics, sports, current events, or your experience or observation.

Appendix B: School Communications

Letter of Invitation

October 4, 2002

Dear High School Principal,

The College Board's recent announcement that a writing test will be included in a new SAT®, now being developed attests to a growing recognition of the importance of good writing skills for success in college and beyond. To help develop the essay portion of the writing test, we are planning an important research study, and your school is one of only 500 nationwide invited to take part. Participation in this study will give your students an opportunity to practice some of the types of writing that are considered to be most important by writing teachers and researchers. Your school does not have to use the SAT to participate.

Participation will require that one or two eleventh-grade English teachers use most of one regular class period, or about 40 minutes, to administer an essay exercise with topics that we will provide. Since we are seeking a cross section of the student population, these classes should not be composed solely of your best English students but should be representative of your school. We will send instructions for administering the exercise. Completed essays will be read independently by two experienced evaluators, and the essays and evaluations will be returned to your teachers for use as a learning tool. (Teachers might ask students to improve their essays based on the evaluations.)

Materials will be sent to participating schools in late October, and schools can administer the exercise anytime before November 21. The completed essays must be shipped back to us no later than November 21. (A return envelope and shipping instructions will be included with the materials.) We will return the evaluations to you, along with the student responses and information comparing your school's performance with that of our national sample, in January. Your school's performance data will be included in the national sample, but it will not be shared in school-identifiable form with any third parties.

If you would like your students to have an opportunity to practice their writing, please respond no later than **October 21** by e-mail to essaystudy@ets.org or by faxing the completed second page of this letter to 609 683-2130 (Attention: Essay Study). Final selection of participating schools will be made by October 25.

If you have questions about the study, please send an e-mail message to the above address or call Regina Mercadante at 609 734-5906. Thank you for your assistance.

SAT Program

Attention: Essay Study

Please respond by October 21, 2002
FAX this form to 609 683-2130
Or E-MAIL all the requested information to essaystudy@ets.org

_____ Yes, our school would like to participate in the essay study.

_____ No, our school cannot participate at this time.

Please include your contact information and e-mail address. We will contact you in late October to let you know if your school has been selected for the study.

School Name _____ School Code _____

School Representative:

Name _____

Address _____

Phone _____ Fax _____

E-mail _____

Numbers of teachers _____, classes _____, and juniors _____ you estimate will participate.

Letter Transmitting Study Materials

Dear Principal:

Thank you for participating in our Essay Research Study. We believe this will be a good learning experience for your teachers and their students as well as being a great help to us.

Enclosed are packets of materials for the English teachers who will be participating in the study. Since the responses to our invitation for help were so overwhelmingly positive, we may not be sending as many test booklets as you requested. Also, we may not be able to score all of the essays that you return. We believe, however, that your teachers will be able to score the additional essays, by using the sample (scored) essays and the scoring directions used by our readers.

Before you distribute the materials to the participating teacher or teachers, please note your school code that is a part of our mailing label. Teachers will need to give this code to their students who will enter it on their test booklets. This will ensure that all the essays written by your students are returned to your school.

If you have any questions, please e-mail (essaystudy@ets.org), FAX (609 683-2130, Attention: Essay Study), or call Regina Mercadante at 609 734-5906.

Thank you again for agreeing to help in this important study.

Sincerely,

SAT Program

Instructions for Administering the Essay Research Study

Before the Administration

1. Check your shipment of materials carefully. Make certain that you have enough essay booklets for the number of students who are participating.

2. Make the school code (AI code) available for your students (it can be found on the mailing envelope sent to your school). Students will need to enter the code on their essay booklets.

3. The essay questions should be inspected only by you or by an instructor who will administer the study exercise.

4. At your discretion, or if students inquire, feel free to explain that these are newly written questions being tried out for possible use in future College Board examinations.

5. When testing is over, collect all test books before you dismiss the students.

6. ***PLEASE RETURN ALL MATERIALS in the FedEx envelope provided by November 21, 2002.***

Administration

1. Announce to students that they will be dismissed after the test is over and all materials are collected.

2. Distribute a booklet to each student. Distribute the booklets from each of the packets in the order that they were received.

3. Instruct students to complete page 1 of the test booklet. Note that the six-digit school code is on the mailing label of the package that your school received. Write this code on the board so that students can enter it in the appropriate space on their test booklets.

4. Allow **25** minutes for writing the essay.

5. Collect a book from each student when the testing is over.

6. You may wish to grade these essays yourself. If you do, please do not put any grade on, or mark any errors on the essays or the test books. It is important that our essay readers/scorers remain unprejudiced in their evaluation of the quality of the answers.

After the Administration/Return of Materials

1. Please ship all of the essays (used and unused) via Federal Express by **November 21, 2002,** using the return prepaid label and envelope provided.

2. Complete the "Cover Sheet for Returning Materials" and return it with your mailing.

If you have any questions or difficulty in returning the materials as instructed, send an e-mail message to essaystudy@ets.org or call Regina Mercadante at 609 734-5906.

Appendix C: Participating Schools

Anacostia High School, Washington, DC

St. Thomas Aquinas High School, Ft. Lauderdale, FL

Bishop Noll Institute, Hammond, IN

Don Bosco Technical Institute, Rosemead, CA

Brighton High School, Brighton, MA

Bronx High School of Science, Bronx, NY

William Cullen Bryant High School, Long Island City, NY

Charlestown High School, Charlestown, MA

Cibola High School, Yuma, AZ

DeWitt Clinton High School, Bronx, NY

Calvin Coolidge Senior High School, Washington, DC

A.J. Dimond Senior High School, Anchorage, AK

Granada Hills High School, Granada Hills, CA

Gwyn Park High School, Brandywine, MD

Harker High School, San Jose, CA

Institute of Notre Dame, Baltimore, MD

John F. Kennedy High School, Fremont, CA

Kaimuki High School, Honolulu, HI

Kaiser High School, Fontana, CA

Kolbe Cathedral High School, Bridgeport, CA

Laora High School, Anaheim, CA

Long Island City High School, Long Island City, NY

Long Reach High School, Columbia, MD

Los Altos High School, Hacienda Heights, CA

Mid-Pacific Institute, Honolulu, HI

Mission San Jose High School, Fremont, CA

Montebello High School, Montebello, CA

Mt. Vernon High School, Mt. Vernon, NY

Daniel Murphy Catholic High School, Los Angeles, CA

Norwalk High, Norwalk, CA

Pomona High School, Arvada, CO

Punahou High School, Honolulu, HI

Ramona High School, Riverside, CA

Ramsay High School, Birmingham, AL

St. Lucy's Priory High School, Glendora, CA

St. Patrick-St. Vincent High School, Vallejo, CA

San Lorenzo High School, San Lorenzo, CA

San Mateo High School, San Mateo, CA

Southfield-Lathrup Senior High School, Lathrup Village, MI

South Gate Senior High School, South Gate, CA

South Pasadena High School, South Pasadena, CA

Takoma Academy, Takoma Park, MD

Townsend Harris High School, Flushing, NY

Walton High School, Bronx, NY

Wayne High School, Ft. Wayne, IN

West Catholic High School, Philadelphia, PA

George Westinghouse Vocational-Technical High School, Brooklyn, NY

Westwood High School, Memphis, TN

Whitney High School, Cerritos, CA

Appendix D: Analysis of Variance Tables

TABLE D-1

Prompt Type X Ethnicity Analysis of Variance with Prompt Topics Nested Within Prompt Types (Analysis for Figure 1)

Source	*df*	*SS*	*MS*	*F*	*p*
Between	15	825.61	55.04	13.06	0.000
Type	1	0.02	0.00	0.00	0.974
Prompt w/Type	2	4.04	2.02	0.48	0.619
Ethnicity	3	776.84	258.95	61.47	0.000
T x E	3	6.21	2.07	0.49	0.688
P x E	6	38.52	6.42	1.52	0.166
Within	2377	10014.06	4.21		
Total	2392	10839.67			

Type X Ethnicity Means and Confidence Intervals

	SAT II		*Persuasive*		*Row*	
	95% Error	*Mean*	*95% Error*	*Mean*	*N*	*Mean*
Asian American	0.234	7.04	0.231	7.02	598	7.03
African American	0.238	5.75	0.236	5.61	578	5.68
Hispanic	0.230	6.14	0.227	6.16	621	6.15
White	0.231	6.93	0.235	7.07	596	7.00
Column	1193	6.46	1200	6.47	2393	6.47

TABLE D-2

Prompt X Ethnicity Analysis of Variance (Analysis for Figure 2)

Source	*df*	*SS*	*MS*	*F*	*p*
Between	15	825.61	55.04	13.06	0.000
Ethnic	3	776.44	258.81	61.43	0.000
Prompt	3	2.78	0.93	0.22	0.883
Ethnic x Prompt	9	44.72	4.97	1.18	0.304
Within	2377	10014.06	4.21		
Total	2392	10839.67			

TABLE D-3

Prompt X Gender Analysis of Variance (Analysis for Figure 3)

Source	*df*	*SS*	*MS*	*F*	*p*
Between	7	181.60	25.94	5.81	0.000
Gender	1	171.62	171.62	38.4	0.000
Prompt	3	6.43	2.14	0.48	0.696
Gender x Prompt	3	6.08	2.03	0.45	0.715
Within	2385	10658.08	4.47		
Total	2392	10839.67			

TABLE D-4

Prompt X Language Analysis of Variance (Analysis for Figure 4)

Source	*df*	*SS*	*MS*	*F*	*p*
Between	10	7.310	0.731	0.16	0.999
Language	2	0.552	0.276	0.06	0.941
Prompt	3	2.525	0.842	0.19	0.907
Language x Prompt	5	1.997	0.399	0.09	0.994
Within	2382	10832.37	4.55		
Total	2392	10839.67			

CollegeBoard

connect to college success
www.collegeboard.com

30481024